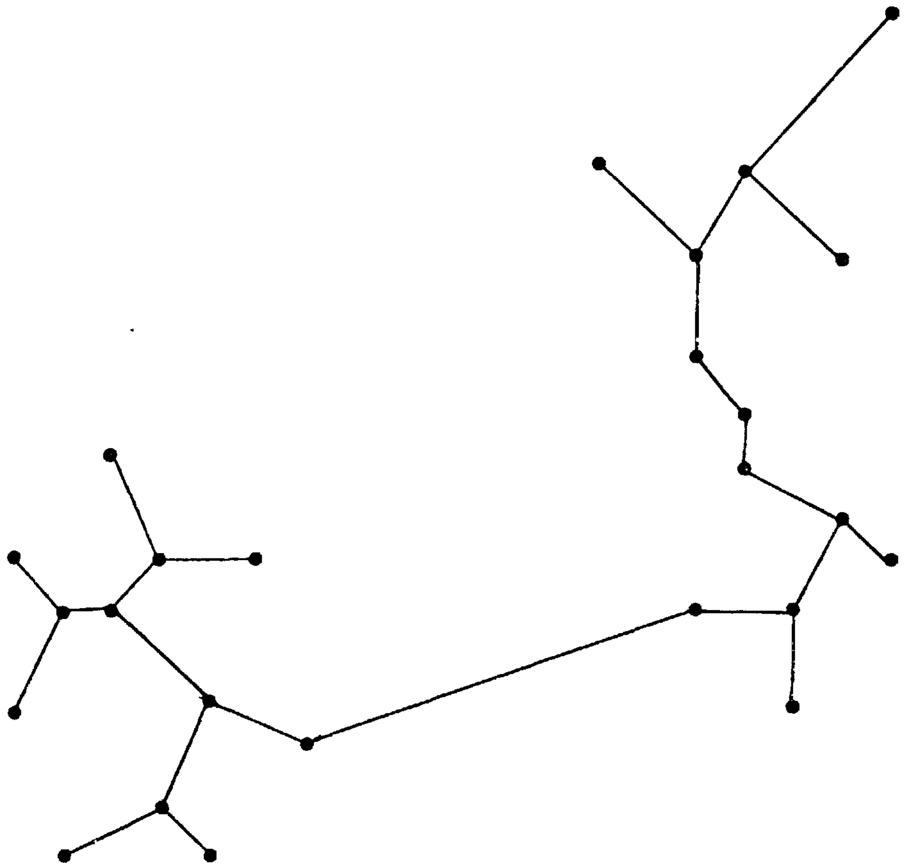
MSTCLUS 1.0.0

PROCEDURE MINIMAL SPANNING TREE

CLUSTER ANALYSIS



by

Chapman P. Gleason

Research Division

Statistical Reporting Service

September 1974

THE MSTCLUS PROCEDURE

The procedure MSTCLUS performs a Minimal Spanning Tree (MST) cluster analysis on a N-dimensional point (node) set (reference 3 and 1.)[1] The Euclidean distance function is used to compute distances between nodes using all variables in the VARIABLES statement. The MST is then partitioned into connected components (clusters) by determining a set of "inconsistent" edges for the MST. An edge is "inconsistent" if its length (euclidean distance between its connecting nodes) is greater than the average edge length of edges in a neighborhood of radius LEVEL DEPTH around each of its connecting nodes plus SIGMA THRESHOLD times the standard deviation of these edges in the neighborhood and if the ratio of the edge length to the average edge length of the neighborhood is greater than the parameter FACTOR THRESHOLD. LEVEL DEPTH, SIGMA THRESHOLD and FACTOR THRESHOLD are parameters which are specified on the PROCEDURE statement. If the MST is to be partitioned more than once using different values of these parameters, then they can be read by using the PARMCARDS statement.

In the language of the educational psychologist, biologist and the sociologist, MSTCLUS perform a single linkage hierarchical cluster analysis. The representation of the MST corresponds to a dendrogram. The distance between nodes is called a similarity measure between individuals. See reference 2 for further discussion.

OUTPUT

MSTCLUS will print the set of nodes and N-dimensional values which are specified on the VARIABLES statement.

The MST arranged by nodes. That is a set of ordered pairs representing

[1] The author of the procedure is Charles Zahn, Stanford University. It was implemented into SAS by Chapman P. Gleason.

the edges connecting each N-dimensional point.

The set of edges joining groups known a priori and the number of edges of this type .

The set of inconsistent edges for the MST.

The set of clusters computed by deleting inconsistent edges in the MST along with the set of nodes in each cluster, and the diameter path in the cluster.

THE PROCEDURE MSTCLUS STATEMENT

The PROCEDURE MSTCLUS statement is of the form:

```
PROC MSTCLUS <NOMST> <NONODES> <READMST> <PUNCHMST>
<LEVELDTH=--> <SIGMATHD=_ _> <FACTORTHD=_ _> <DATA=data_set_name>;
```

Options and Parameters

NOMST When this option is specified the Minimal Spanning Tree will not be printed.

NONODES The option informs the procedure not to print the nodes.

READMST
R This specifies that the Minimal Spanning Tree is to read from a file with ddname 'READMST'. This option should be used after option PUNCHMST has been specified to save the MST.

PUNCHMST
P This option will punch the edges of the MST out on SAS file FT02F001. This saves recomputing the MST in later analysis of the same data set. File FT02F001 can be overridden with JCL to write the data set on tape or disk.

LEVELDTH=_ _
LD=_ _ This acronym for the parameter LEVEL DEPTH specifies the subtree depth of a neighborhood to be taken by the program around each node (point) to determine edge inconsistency. The default value is 2.

SIGMATHD=_ _
ST=_ _ SIGMA THRESHOLD, this value along with the value of FACTORTHD will be used to partition the MST into clusters. The default value is 3.

FACTORTH=_ _
FT=_ _ FACTOR THRESHOLD, this value and the SIGMATHD value are used as the criteria to partition the MST into clusters. The default value is 2.

PROCEDURE INFORMATION STATEMENTS

VARIABLES statement

The variables specified in the VARIABLES statement are used to compute the distances between nodes. A VARIABLES statement must accompany the procedure MSTCLUS when there are character variables in the SAS data set not mentioned on the CLASSES statement, an attempt to compute distance with a character variable results in a fatal error. This statement is of the form:

```
VARIABLES
VAR          variable_name_1 <variable_name_2 ... variable_name_k>;
```

CLASSES statement

The Classes statement is used to inform the procedure of any a priori knowledge of group membership of the nodes. For example, we might know from previous analysis that certain nodes belong to certain groups and want to see if these groups are in fact distinct in the MST, or possibly if there are subgroups or subclusters of these well defined groups. This variable must be a character variable. The CLASSES statement for this procedure is of the form:

```
CLASSES
CLASS        variable_name;
```

PARMCARDS statement

When the MST is to be partitioned into clusters more than once, the values of the clustering criteria (LEVEL DEPTH, SIGMA THRESHOLD, FACTOR THRESHOLD) can be read from PARMCARDS. This saves computation time since the data does not have to be transferred into the procedure again and the MST does not have to be recomputed. The parameter LEVEL DEPTH is the first value, SIGMA THRESHOLD is the second, and FACTOR THRESHOLD is third. At least one blank must separate each value on a card.

EXAMPLE:

In the following example the MST is computed and printed for a two dimensional node set. The default values are taken for the clustering parameters.

```
DATA EXAMPLE;

INPUT CLASS $ 1-2 X1 5-10 X2 12-16;
CARDS;

SW      2    3
SW      4    4
SW      7    5
SW      5    3
SW      5    6
NW     16   17
NW     15   15
NW     13   17
NW     18   15
NW     15   13
NW     18   10
NW     19    9
NW     17    8
NW     17    6
NW     15    8
NW     19   20
NW     16   12
NW     16   11
SW      3    8
SW      4    9
SW      2    8
SW      6    9
SW      3   11
SW      1    9
SW      1    6

PROC MSTCLUS; VAR X1 X2; CLASSES CLASS; TITLE  'ZAHN SAMPLE DATA';
```

1. Anderberg Michael R., Cluster Analysis for Applications. New York: Academic Press, 1973.

2. Gower J. C. and Ross G. J. S., "Minimal Spanning Tree and Single Linkage Cluster Analysis," Applied Statistics, Vol. 18 (1969), No. 1, pp. 54-64.

3. Zahn, C. T., "Graph Theoretical Methods for Desecting and Describing Gestalt Clusters." IEEE Transactions on Computers, Vol. C-20, No. 1, January, 1971, pp. 68-86.